We Are All Data Now

Editorial by Ali Krzton, Research Data Management Librarian, Auburn University

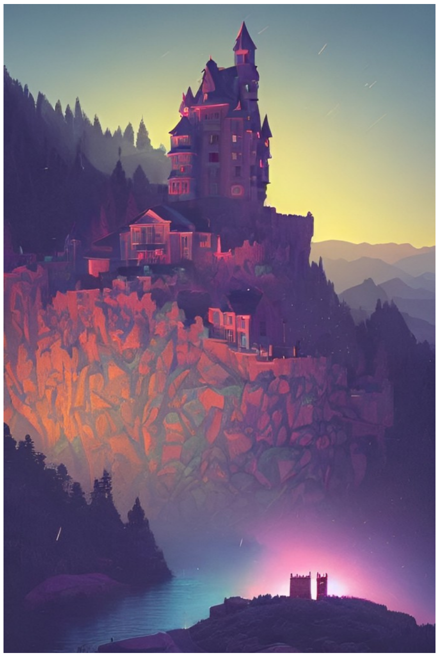

Image generated with the Wonder app on iOS by Ali Krzton

An abiding concern for the responsible use of data is nothing new for those of us among the ranks of data professionals. Whether from the perspective of research, education, commerce, or policy, careful consideration of data ethics is an important aspect of our work. The recent increase in the accessibility and prevalence of “generative AI” technology adds a new dimension to the digital landscape and a new level of difficulty for would-be facilitators of information literacy. Briefly, generative AI uses complex, impenetrable statistical models to produce novel combinations of text or images in response to user prompts. The best-known example may be ChatGPT, which can in mere seconds write a poem in the style of Edgar Allen Poe, a tailored marketing pitch, or a research paper that may or may not be accurate. What does this mean for us?

As always, the story begins with the data. Rapid advances in generative AI tools such as ChatGPT depend on access to a truly massive corpus of data to scale up their learning. Thus, it is no surprise that material scraped and aggregated from the open web for free would be a critical source of training data. However, the people who created and published the articles, blogs, and forum posts used to shape large language models like those that power ChatGPT could not have foreseen, much less intended, their work would be used this way. These concerns are now echoed by a broader community of content creators including authors, artists, and coders. Recently, some fanfiction writers expressed dismay at compelling evidence that the Archive of Our Own repository was used to train the GPT-3 LLM [1]. This example, while seemingly obscure, shows that nearly anything on the web might end up fueling the next AI iteration.

Researchers skeptical of open data sharing have long objected that putting their data online without restriction could allow bad actors to misuse it, or at the very least avoid attributing the source. The increasing emphasis on machine-readability and automated processing of research data [2] makes the attribution problem even more acute. In response to a lawsuit brought by developers of open source code on GitHub alleging that GitHub's AI-powered Copilot tool violated their licenses, Microsoft and OpenAI defense relies in part on the lack of particular examples of misappropriated code [3].

In fact, it would be impossible for anyone to say which specific passage, image, script, or dataset was used to produce any given generative AI output by design. These models can only be built by digesting inputs down to vectors that do not translate to human-intelligible dimensions. This is how things produced by AI can be at once novel and derivative. Technologist Jaron Lanier warned of the attitude that such developments can only improve human understanding in a prescient passage from 2010: "A fashionable idea in technical circles is that quantity not only turns to quality at some extreme of scale, but also does so according to principles we already understand. Some of my colleagues think a million, or perhaps a billion, fragmentary insults will eventually yield wisdom that surpasses that of any well-thought-out essay, so long as sophisticated secret statistical algorithms recombine the fragments [4]." For those who hold improvements in AI capability as an end in itself, the necessary stripping away of all context is an acceptable cost.

What are we to make of this massive disruption in our ability to control digital representations created either by us or of us? Continuing to learn everything we can about it is essential. Sharing our knowledge of how AI systems work with those around us is crucial so that we can all curate our personal data more intentionally. Helping more people be aware of possible interactions between AI and what they publish (or perhaps even enter) online could have a greater impact than any of us currently appreciate.

References

[1] Eveleth, R. (15 May, 2023). The fanfic sex trope that caught a plundering AI red-handed. *Wired.* https://www.wired.com/story/fanfiction-omegaverse-sex-trope-artificial-intelligence-knotting/

[2] Huerta, E. (14 February, 2022). FAIR and AI-ready scientific datasets. *SpringerNature Research Data Community.* https://researchdata.springernature.com/posts/fair-and-ai-ready-scientific-datasets

[3] Brittain, B. (27 January, 2023). OpenAI, Microsoft want court to toss lawsuit accusing them of abusing open-source code. *Reuters.* https://www.reuters.com/legal/litigation/openai-microsoft-want-court-toss-lawsuit-accusing-them-abusing-open-source-code-2023-01-27/

[4] Lanier, J. (2010). *You Are Not a Gadget.* New York: Vintage Books, p. 49.